CISC 271, Winter 2021
Assignment #2: Regression and Cross-Validation
Due by 11:30AM on Friday, February 26, 2021

The subject matter for this assignment is data of the environment, specifically air quality. The data were downloaded from the University of California at Irvine, which maintains extensive data sets for machine learning. The data were gently processed for use in this class.

Coding for this requires multiple uses of linear regression, which is the solution of an over-determined linear equation. You may use the `MATLAB` builtin function `linsolve`, or the "back-slash" operator \, or any builtin function. You may *not* use functions for cross-validation, or any other functions from a `MATLAB` toolbox, because you are expected to code these by yourself.

The technical problem in linear regression, for this assignment, is to estimate a weight vector $\vec{w}$ for a design matrix A that is "tall thin", and a data vector $\vec{c}$. The vector $\vec{w}$ is an approximate solution to the regression problem

$$A\vec{w} \approx \vec{c}$$

Please read the details and instructions carefully before you begin to work on the problem. The second question in this assignment is modestly difficult because it is intended to be a practical introduction to a method of evaluating algorithms for linear regression. There must be a single results section and a single discussion section on your report. The results section of the report must contain two tables and up to one figure; more or fewer tables or figures, may produce deductions from your grade on this assignment.

Statement of Academic Integrity

Sharing of all or part of a solution to this assignment, whether as code or as a report, will be interpreted as a departure from academic integrity. This includes sharing of the assignment after the due date and after completion of this course.

Learning Outcomes

On completion, a successful student will be able to:

- Import data from a Comma Separated Values (CSV) file into `MATLAB`
- Develop and implement a method for managing missing data
- Standardize data for use in linear regression
- Compare different regressions in a consistent manner
- Implement a 5-fold cross-validation of linear regression
- Evaluate the results of a cross-validation

Preliminary: The Data

This assignment uses data that the instructor gathered from the Machine Learning Repository of the University of California at Irvine. The original data, and a full description, can be found at

`https://archive.ics.uci.edu/ml/datasets/air+quality`

Briefly, the original data had 15 variables. The instructor removed 4 variables: date of readings, time of readings, relative humidity of readings, and absolute humidity of readings. The first two variables are not needed and the last two variables are highly correlated with temperature, for complicated reasons having to do with the locale of the sensors. From the website:

> *The data set contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer . . . Missing values are tagged with -200 value . . .*

This is "real" data so not all values were available when the data were gathered. In the introduction section of your report, you must describe and verbally defend your choice for how to manage the missing data. You are expected to explore beyond the course notes and textbook in making this choice.

The data are in the file `airquality.csv` that is in the ZIP file for this assignment. The data can be loaded using the `csvread` function that is provided in `MATLAB`. You will need to "skip" the first row when reading the data for numerical processing. For your report, you will need to understand the names of the variables in the dataset.

Preliminary: The Code

The starter code is structured as three functions. The base function, which is the function that will be called when the graders invoke your file name, is `a2`.

You can, if you wish, change the name of this function to match the name of your submission file. `MATLAB` will execute the code correctly even if the function name and the file name differ. The base function will then invoke two functions in succession.

The base function will invoke the code for the first question. This first code will return two variables: the RMS errors of linear regression and the index of the lowest RMS error. The base function will then invoke the code for the second question. This invocation uses the index from the first question and computes the 5-fold cross-validation for the corresponding variable.

The base function will return four values. The TAs will examine these values and use them as part of the grade for your assignment.

DO NOT MODIFY THE BASE FUNCTION. MODIFY ONLY THESE FUNCTIONS:

`a2q1` `a2q2` `mykfold (optional)`

Question 1: Variable of Best Regression — 10% of Final Grade

For this question you will need to modify the function `a2q1` in the starter code.

The technical problem for this question is to find the variable in the data that is best explained by the other variables. You are expected to do this by selecting the variable that, when chosen as the $\vec{c}$ vector, has the lowest RMS error of fit when the remaining variables are gathered into a data matrix A. In this document, the data are assumed to have been converted into a data matrix `Amat` that has the chemical readings as variables and the air-quality data as observations.

You should use every available observation of the data to search for the regression variable. You should describe what, if any, data standardization you used. You should also describe and justify whether or not you used an intercept term in your linear regression.

The starter code for this assignment will return a row of small numbers; your completed code should return the RMS error for the regression of each variable in terms of all of the other variables.

For example, if there were three variables in the data set, the function `a2q1(0.0000)` would return a matrix such as

```
[0.1 0.2 0.3]
```

The starter code will also return the index of the variable with the smallest RMS error; for the above example, this would be

```
1
```

You may report RMS errors either in the units of the relevant chemical variable or in standardized units, depending on your choice of whether or not to use data standardization.

In your report, the values of the RMS errors must be presented in Table 1. The caption of the table must state the index of the variable that is best explained by the other variables *and* the corresponding name of the chemical variable. You can, optionally, plot the dependent variable and the regression. An example of the optional figure for this assignment is Figure 1.

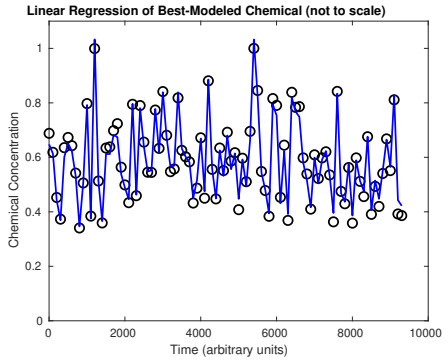

Figure 1: Example of an optional plot of data and a linear regression to these as dependent data. For this plot, the chemical concentrations were scaled to a maximum of 1.

1.1: Linear Regression

Each column of `Amat` will need to be considered as a dependent variable, in this document referred to as `cvec`. The other columns of `Amat` will need to be considered as the independent variables, in this document referred to as `Xmat`. You may wish to standardize the data. If you choose to use an intercept term, you will need to augment `Xmat` with a column that is the $\vec{1}$ vector.

The RMS error of regression of each `cvec` must be computed and returned as the corresponding entry of the variable `rmsvars` of the function `a2q1`. The index of the smallest RMS error must be computed and returned as the variable `lowndx` of the function `a2q1`.

1.2: Methods

Your methods must include a narrative description of your computation. For example, in finding the best data variable, do not simply say something like "I used a `for` loop". You should give a reader the logic that underlies your code, not a line-by-line description of its implementation.

1.3: Discussing The Results

Your discussion should compare and contrast your numerical results and any observation you have on, for example, the merits of using the second-best chemical variable as the dependent variable. You should describe any other effects related to the choices you made in your implementation.

For this assignment, you are encouraged to provide a modest amount of creativity – it is a good idea to try to engage the readers of your report.

Question 2: Cross Validation of Regression 10% of Final Grade

The problem for this question is to determine the reliability of choosing one chemical variable to act as a proxy for the other chemical variables in the given data. This reliability will be found by performing a 5-fold cross-validation of the given data.

Important: RMS errors must be reported in the units of the chemical that is the dependent variable. If you standardized the data for linear regression, the standardization may need to be inverted, or recomputed with an intercept term, to find the correct units.

You can select the five folds of data by a method of your choice. In the introduction of your report, you must state your choice and verbally defend the choice. The instructor's notes recommend a random selection, which requires a form of indirect indexing that is more complex than the indexing that was used in previous assignments.

For each fold, you must first use 4/5 of the data to "train" your regression. This training is the computation of the weight vector of a linear regression.

For each fold, you must then use 1/5 of the data to "test" your regression. This testing is the application of the weight vector from the training phase; you must use the same choices of standardization and an intercept term that you used in the training phase.

For each fold, the RMS error of training must be computed and returned as the corresponding entry of the variable `rmstrain` of the function `a2q2`. For each fold, the RMS error of testing must be computed and returned as the corresponding entry of the variable `rmstest` of the function `a2q2`.

In your report, the values of the RMS errors of the folds must be presented in Table 2. The caption of the table must state the index of the variable that is best explained by the other variables *and* the corresponding name of the chemical. You can, optionally, summarize the RMS errors of testing and the RMS errors of testing as two overall values.

2.1: Loading The Data

The data must be loaded exactly as they were loaded for Question 1 of this assignment. You must use the column specified as the value of the input variable as the dependent vector $\vec{c}$, and the remaining data as the data matrix A.

2.2: Fold Selection

There are many ways to select folds, including algorithmic selection and random selection. You must describe and verbally defend your choice. Some choices require more elaborate coding than others and some choices introduce biases into the results, so this choice has consequences.

The starter code includes the `MATLAB` function `mykfold`, which is the instructor's starter code for randomly selecting the folds. Because you do not need to use this code, this code is not called from the other starter code.

3: Grading Guide

We will test your code by invoking the function that you uploaded. Your grade will be reduced if: you plot more or fewer than the specified number of figures; your code outputs anything other than the specified values; or you otherwise deviate in your implementation from these specifications.

The TA's have been instructed to use this guide when they mark your assignment. Your grade will be based on the numerical results and on the report. The distribution of points for the assignment grade are:

$6/40$ points: all and only the numerical values that are produced by the code and that are presented in the results

$9/40$ points: quality of the code in the modified "starter" functions, and any other changes in the submission file that was used to generate values and plots for the report

$25/40$ points: quality of the report, especially including the figures and descriptions; clarity may be assessed, in part, by the written introduction, verbal defense of choices, and the discussion of results

What to turn in:

- You will submit your answers electronically as two files. The code will be tested by one or more graders. The PDF report will be read by one or more graders and will be checked, using electronic methods, to ensure that it meets professional standards for originality.
- The code must be in one `MATLAB` file (`a2_xxxxxxxx.m`). This file will contain all of the code needed to verify that the values and tables in the report can be reproduced. The functions must produce the values for your tables and the figure.
- Your functions must take no arguments, return the specified values, and require no user input or action such as using the "enter" key. Running this function should produce, on the console, every value that is in the report; the function should also produce any plot that is in your report. The function should produce no other values or figures. The graders will compare your computed values to the values in the report and may deduct marks from the report for differences between any reported value/plot and the corresponding computed value/plot.
- The report must be in a single PDF file (`a2_xxxxxxxx.pdf`). The PDF file must include a description of how you tested your code. You can also include notes, comments on the problems, and assumptions you have made, as appropriate.
- The assignment must be submitted using the Queen's "onQ" software.

Grading Considerations:

- The quality of your report will be considered. You need, at minimum, to conform to the "student version" of the report style in the onQ website; you may wish to consider the "grader version" that we will use for assessing your report.

- The quality of your `MATLAB` code will be considered. Your code should be appropriately indented, sufficiently commented, and otherwise be appropriate software.
- The output of your code will be considered.
- Your code can use functions provided by `MATLAB`, but the code that you submit *must* be your original work. You may not use any builtin functions that perform k-fold cross-validation.
- Code that causes `MATLAB` to produce an error or warning will result in a failing grade.
- You may assume that the file `airquality.csv` is in the current directory when a grader tests your code.

Policies:

- You must complete these questions individually.
- Although you are allowed to discuss the questions with other students, you must write your own answers and `MATLAB` code.
- The syllabus standards apply to this assignment.
- Lateness policy applies starting the minute after the submission deadline, at a rate of 20% off the assignment value per calendar day. *Please note: the time in the onQ system is beyond your control, so submitting within an hour of the deadline is inherently a risky process for which you assume full responsibility.*